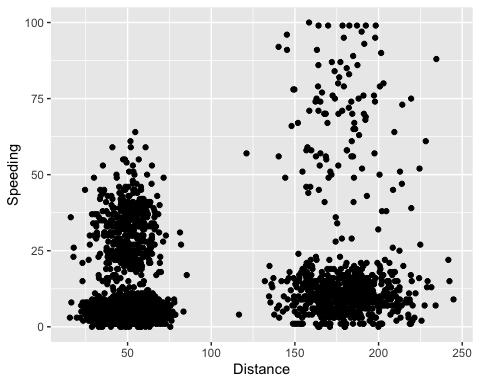
# BAN502 Module 6 - Assignment 1

## Amanda Odum

#For this assignment you will need the following libraries: tidyverse and tidymodels. Before beginning the assignment tasks, you should read-in the “trucks.csv” dataset into a data frame called “trucks”. In this dataset, Driver\_ID is a unique identifer for each delivery driver, Distance is the average mileage driven by each driver in a day, and Speeding is the percentage of the driver’s time in which he is driving at least 5 miles per hour over the speed limit.  
  
library(tidyverse)  
library(tidymodels)  
trucks <- read\_csv("trucks.csv")

#Task 1: Plot the relationship between Distance and Speeding. Describe this relationship. Does there appear to be any natural clustering of drivers?  
  
ggplot(trucks,aes(x=Distance, y=Speeding)) + geom\_point()



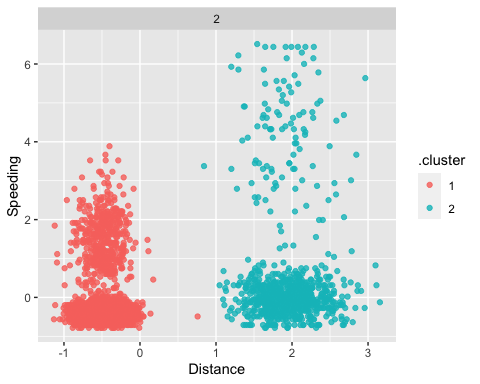
*It appears that drivers that spend a greater percentage of their drive speeding tend to also average higher daily distances Natural clusters appear between 0 and 25 percent of time spent speeding for both 0-75 miles (avg) per day and 150-225 miles (avg) per day.*

#Task 2: As we did in the second clustering example, create a new data frame called “trucks\_cleaned” that contains the scaled and centered variables. Two notes: 1) The “predictor” variables in the recipe are “Distance” and “Speeding” and 2) There is no need to create dummy variables as there are no categorical variables in the data.  
kmeans\_recipe = recipe(~ Distance + Speeding, trucks)   
trucks\_dummy = kmeans\_recipe %>%   
 step\_scale(all\_numeric()) %>%  
 step\_center(all\_numeric())   
  
trucks\_dummy = prep(trucks\_dummy, trucks) #prepares the recipe  
trucks\_cleaned = bake(trucks\_dummy, trucks)  
  
#summary(trucks\_cleaned)  
#str(trucks\_cleaned)

#Task 3 Use k-Means clustering with two clusters (k=2) to cluster the “trucks\_cleaned” data frame. Use a random number seed of 64. Use augment to add the resulting clusters object to the the “trucks” data frame. Design an appropriate visualization to visualize the clusters. Comment on the clusters.  
  
set.seed(64)  
clusts =   
 tibble(k = 2) %>%  
 mutate(  
 kclust = map(k, ~kmeans(trucks\_cleaned, .x)),  
 tidied = map(kclust, tidy),  
 glanced = map(kclust, glance),  
 augmented = map(kclust, augment, trucks\_cleaned)  
 )  
  
clusts

## # A tibble: 1 × 5  
## k kclust tidied glanced augmented   
## <dbl> <list> <list> <list> <list>   
## 1 2 <kmeans> <tibble [2 × 5]> <tibble [1 × 4]> <tibble [4,000 × 3]>

clusters =   
 clusts %>%  
 unnest(cols = c(tidied))  
  
assignments =   
 clusts %>%   
 unnest(cols = c(augmented))  
  
clusterings =   
 clusts %>%  
 unnest(cols = c(glanced))  
  
set.seed(64)  
clusters = kmeans(trucks\_cleaned, 2)  
  
trucks = augment(clusters, trucks)  
#str(trucks)  
  
p1 =   
 ggplot(assignments, aes(x = Distance, y = Speeding)) +  
 geom\_point(aes(color = .cluster), alpha = 0.8) +   
 facet\_wrap(~ k)  
p1

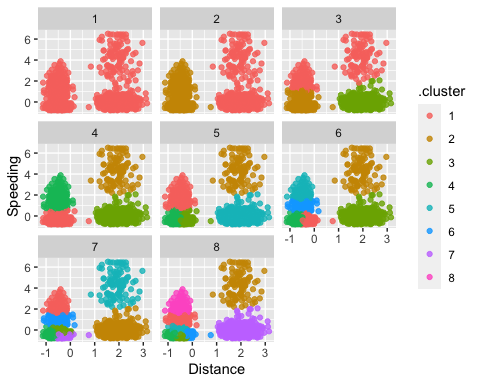


*Cluster 1 has a lower average distance traveled and lower percentage of speeders overall, whereas Cluster 2 has a greater variance of speeding and higher average distance traveled. It appears that distance traveled and percentage of travel time spent speeding are positively correlated.*

#Task 4: Create a visualization to show how the cluster appear from values of k from 1 to 8. Use a random number seed of 412. Which value of k appears to be most appropriate for this data?  
  
set.seed(412)  
clusts =   
 tibble(k = 1:8) %>%  
 mutate(  
 kclust = map(k, ~kmeans(trucks\_cleaned, .x)),  
 tidied = map(kclust, tidy),  
 glanced = map(kclust, glance),  
 augmented = map(kclust, augment, trucks\_cleaned)  
 )  
  
clusts

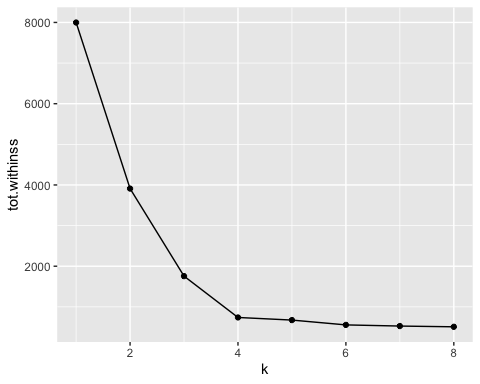
## # A tibble: 8 × 5  
## k kclust tidied glanced augmented   
## <int> <list> <list> <list> <list>   
## 1 1 <kmeans> <tibble [1 × 5]> <tibble [1 × 4]> <tibble [4,000 × 3]>  
## 2 2 <kmeans> <tibble [2 × 5]> <tibble [1 × 4]> <tibble [4,000 × 3]>  
## 3 3 <kmeans> <tibble [3 × 5]> <tibble [1 × 4]> <tibble [4,000 × 3]>  
## 4 4 <kmeans> <tibble [4 × 5]> <tibble [1 × 4]> <tibble [4,000 × 3]>  
## 5 5 <kmeans> <tibble [5 × 5]> <tibble [1 × 4]> <tibble [4,000 × 3]>  
## 6 6 <kmeans> <tibble [6 × 5]> <tibble [1 × 4]> <tibble [4,000 × 3]>  
## 7 7 <kmeans> <tibble [7 × 5]> <tibble [1 × 4]> <tibble [4,000 × 3]>  
## 8 8 <kmeans> <tibble [8 × 5]> <tibble [1 × 4]> <tibble [4,000 × 3]>

clusters =   
 clusts %>%  
 unnest(cols = c(tidied))  
  
assignments =   
 clusts %>%   
 unnest(cols = c(augmented))  
  
clusterings =   
 clusts %>%  
 unnest(cols = c(glanced))  
  
p2 =   
 ggplot(assignments, aes(x = Distance, y = Speeding)) +  
 geom\_point(aes(color = .cluster), alpha = 0.8) +   
 facet\_wrap(~ k)  
p2



*The most appropriate value appears to be k = 4.*

#Task 5: Create a plot of k versus within cluster sum of squares. Hint: We did this in the first clustering lecture. Which value of k appears to be best?  
  
ggplot(clusterings, aes(k, tot.withinss)) +  
 geom\_line() +  
 geom\_point()

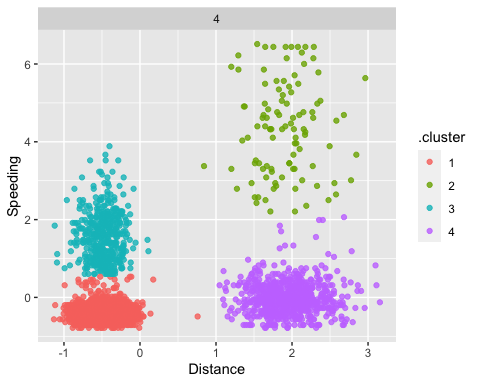


*It appears that k = 4 is the best value.*

#Task 6: Repeat Task 3 for the number of clusters that you identifed in Task 5. Use the same random number seed as in Task 3. Don’t forget to include your visualization. Comment on the resulting cluster.  
  
set.seed(64)  
clusts =   
 tibble(k = 4) %>%  
 mutate(  
 kclust = map(k, ~kmeans(trucks\_cleaned, .x)),  
 tidied = map(kclust, tidy),  
 glanced = map(kclust, glance),  
 augmented = map(kclust, augment, trucks\_cleaned)  
 )  
  
clusts

## # A tibble: 1 × 5  
## k kclust tidied glanced augmented   
## <dbl> <list> <list> <list> <list>   
## 1 4 <kmeans> <tibble [4 × 5]> <tibble [1 × 4]> <tibble [4,000 × 3]>

clusters =   
 clusts %>%  
 unnest(cols = c(tidied))  
  
assignments =   
 clusts %>%   
 unnest(cols = c(augmented))  
  
clusterings =   
 clusts %>%  
 unnest(cols = c(glanced))  
  
set.seed(64)  
clusters = kmeans(trucks\_cleaned, 4)  
  
p3 =   
 ggplot(assignments, aes(x = Distance, y = Speeding)) +  
 geom\_point(aes(color = .cluster), alpha = 0.8) +   
 facet\_wrap(~ k)  
p3



*By splitting the data into four clusters, we have two groups of significant speeders and two of low to non-speeders (one traveling shorter distances and one traveling longer distances in each category). With this visualization, it is clear that there is a higher concentration of shorter distance drivers that are “part-time” speeders, whereas the longer distance group has all of the “full-time” speeders.*